学校编码：10384　　分类号　　密级
学号：　　UDC

厦门大学

硕士学　位　论　文

不平衡数据分类方法研究

Research on Imbalanced Data Classification Problem

指导教师姓名：
专　业名　称：
论文提交日期：
论文答辩时间：
学位授予日期：

答辩委员会主席：_____
评　　阅　　人：_____

20　年　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“ √ ”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

不平衡数据分类方法研究是机器学习和数据挖掘领域的一个研究热点。在各个学科的应用领域中，数据不平衡的现象很常见，但传统分类器在解决不平衡数据分类问题时，往往出现分类器性能大幅度下降的情况。自 21 世纪初开始，国内外对不平衡数据分类的研究越来越重视，国际上的学者不断在相关会议上对这一课题进行深入探讨。尽管针对不平衡数据分类问题已经取得了很多研究成果，但仍有很多值得研究的问题。

本文主要对不平衡数据分类问题进行深入研究，通过分析不平衡数据影响分类器性能的主要因素，总结目前解决不平衡数据分类问题的主要方法，进而在基于采样和基于集成分类器两方面对不平衡数据分类性能的提升进行了研究。在基于采样的不平衡数据分类方法上，提出了基于 FarthestFirst 的新聚类降采样算法、小样本加权重随机抽样算法，获得了较佳的分类效果；在基于集成分类器的不平衡数据分类方法上，分别基于单个基分类器和基于多个基分类器提出了多种改进的算法，通过实验对比，验证了本文提出的改进的集成分类算法均能有效提高分类器的性能，得出了结合数据预处理及分类器集成方法才能更高效地提升分类器性能的结论，同时总结出了不同类型数据集适用的改进的集成分类算法。利用总结出的结论在生物信息学上进行应用，根据生物信息学不同数据不同的特点，选择合适的改进的分类算法，通过实验表明，改进的分类算法能有效提升分类器在生物信息上的分类性能，可以帮助降低实验验证成本，有较大的实际应用价值。

关键词：类别不平衡；分类；集成学习

ABSTRACT

Research on imbalanced data classification problem is a hot topic in the field of machine learning and data mining. In the application of various disciplines, class imbalance phenomenon is very common. However, the traditional classifiers behaved badly when they were used to solve imbalanced data classification problems. From the beginning of the twenty-first century, more and more researchers, both at home and abroad, pay attention to the study of imbalanced data classification problem. International scholars continuously explore this topic at the relevant meetings. Despite the imbalanced data classification problem has made many achievements, there are still many problems worth studying.

In this paper we do research on the imbalanced data classification problem. We analyze the main factors of affecting the imbalanced data classification performance, summarize the current main methods to solve the imbalanced data classification problems, and then try to enhance the performance of classification based on the sampling method and ensemble classification method. Based on the sampling method, we employ two sampling methods for the imbalanced data classification. One is under-sampling by FarthestFirst clustering; the other is weighted random sampling. Both of them obtain better performance. Based on ensemble classification method, we propose several improved classification methods, which based on single base classifier or based on a number of base classifiers. Experiments verify the ensemble classification algorithm proposed by this paper can effectively improve the performance of classifier, show the combination preprocessing and ensemble classification method can more effectively improve the performance of classifier. In the same time, we can summarize the conclusion of the different improved classification methods for different types of data sets. Using the conclusion in bioinformatics applications, according to the characteristics of different bioinformatics data, we can choose appropriate improved classification method.

Experiments show that the improved classification methods can effectively improve the performance of classifier in the biological information. With the help of reducing the experiment cost, these improved classification methods have great practical application value.

Key Words: class imbalanced; classification; ensemble learning

目录

Contents

第一章 绪论

1.1 课题背景及意义

当今是一个数据爆炸的时代，数据量已从 TB 级别上升到 PB 乃至 EB 级别，面对海量的数据，如何进行数据挖掘，从中获取有用的信息就显得尤为重要。数据挖掘有很多研究课题，分类问题是其中一个重要的研究课题。分类（Classification）是指从数据中选出已经分好类的训练集，在该训练集上运用数据挖掘分类技术，进行分析和学习，发现隐藏在数据内的规律，建立分类模型，从而能对未知的测试样本进行分类预测。目前对传统的分类问题已经有很多成熟的算法，如 K 近邻算法、决策树算法、人工神经网络算法、贝叶斯算法、支持向量机算法等，它们在数据挖掘的很多领域中都有应用，而且也都获得了很好的分类效果。

虽然传统的分类算法可以取得较好的分类效果，但是它们大都是建立在数据集分布均衡的前提下的，即数据集中各类样本的数目大体一致。可是在各个学科的应用领域中，更常见的是不平衡的数据集。对于二类问题来说，一般都是一个类别的样本数目比另一个类别的样本数目大很多，其中样本数目多的类称为大类，样本数目少的类称为小类。比如在金融欺诈检测[1]中，一般而言，大多数客户的交易行为都是正常的，只有极个别客户可能是潜在的欺诈行为，可能 10 万笔交易中才存在 1 笔欺诈行为；此外在医疗诊断[2]、网络入侵检测[3]、反垃圾邮件[4]、石油勘探[5]等领域，也都存在数据集不平衡的问题。在这些领域中，有些数据不平衡问题是固有的，因为本身小类样本发生的概率就低。还有一部分是因为小类样本需要来自实验验证，而大类样本不需要实验验证，因此获取大类样本成本低，小类样本成本高，从而导致数据集中出现大类远远多于小类的情况。

由于传统分类算法总是以分类模型的总平均分类精度最大为训练目标，不考虑每个类的相对分布情况，当用传统的分类器解决不平衡数据分类问题时，往往出现分类器性能大幅度下降的现象，得到的分类器具有很大的偏向性，分类器倾向于大类，本属于小类的样本往往被错分到大类。这样的分类器在小类上的效果

很差。但实际问题中却常要求小类的检测率足够高，大类的错误率足够小，因为小类样本通常比大类样本重要得多。同样是金融欺诈检测问题，传统的分类器很容易将欺诈行为也分类为正常行为，但是把欺诈行为当作正常行为对银行造成的损失的代价往往比把正常行为误当作欺诈行为的代价高得多。而在医疗诊断上，如果把病人误诊为正常人，耽误了最佳治疗时间，造成的损失更是不可估量。所以，对不平衡数据的正确分类成了亟待解决的问题。对不平衡数据分类问题的研究具有很强的理论意义和应用价值，针对这一问题的研究成果可以大大推动机器学习方法在实际工程中应用的步伐，扩大应用范围，真正为工业生产和社会经济带来较大的经济效益[6]。

自 21 世纪初开始，国内外对不平衡数据集分类的研究越来越重视，国际上的学者不断在相关会议上对这一课题进行深入探讨，2000 年人工智能领域的学术会议 AAAI[7]，2003 年机器学习领域的会议 ICML[8]，都特别针对不平衡数据集的学习问题召开了专题讨论会，2004 年美国计算机协会(ACM)也针对这一专题出版了一期通讯[9]，这些高级别会议，从机器学习框架、算法等方面产生了很多新的研究成果，并且在实际应用中取得了较好的效果。不平衡数据分类问题已经成为了机器学习和数据挖掘领域的一个研究热点。

1.2 国内外研究现状

随着越来越多的国内外学者加入不平衡数据分类问题的研究，影响分类器性能的因素研究方面、算法的改进方面等都取得了很大进展。很多学者根据影响分类器性能的因素对算法进行了有针对性的改进，还有一些学者通过有机集成各种方法，多学科互相渗透，协同解决实际应用中的复杂问题，在各领域进行了有针对性的应用。

在对不平衡数据集分类器性能的影响因素研究方面，Japkowicz N 和 Stephen S 对不平衡数据问题进行了系统研究，分别从概念复杂度、训练样本规模、类间不平衡程度等方面进行实验研究，试图找出数据对分类器性能影响的原因。研究发现当概念复杂度较低时，类间不平衡程度对分类器性能产生的影响不大，提高训练样本规模可以减少不平衡数据对分类器性能的影响，不同的不平衡数据分布

特征对不同的分类方法影响程度有差异[10]。Batista G E 和 Prati R C 等人通过多组对比实验，研究类间和类内不平衡性在影响分类器性能方面的程度差异，发现分类器性能下降有时不完全是不平衡数据引起，还可能是受多类之间重叠样本的影响[11]。Zhou Z H 和 Zhang M L 通过实验证明了训练集不平衡程度、学习任务的复杂程度、训练集规模和分类算法是影响分类器性能的主要因素[12]。

针对训练集样本规模对不平衡数据分类器性能影响，国内外学者提出了重构数据集的思想，进而提出了多种改进的采样算法。Chawla N V 和 Bowyer K W 提出了虚拟少数类过采样算法 SMOTE(synthetic minority over-sampling technique)，通过生成人工样本来对小类样本向上采样[13]。de la Calleja J 和 Fuentes O 提出基于空间距离的过采样方法，并通过局部线性加权回归对实现分类算法[14]。林舒杨和李翠华等人提出了一种降采样方法，利用 K-means 方法对大类样本进行聚类，然后提取聚类中心，以达到降采样的目的[15]。

更多的学者则主要对分类算法进行改进。根据现有分类算法如代价敏感学习、支持向量机算法、决策树算法、神经网络算法在解决不平衡问题时的不足，对算法进行改进，通过调节各个类别之间的代价函数、引入惩罚因子等有利于小类的措施，或者算法间取长补短进行集成，进一步提高了分类算法在不平衡数据集上的分类性能。

1、基于组合集成方法的分类器可以改进弱分类算法的性能，Boosting 方法就是其中一种，但是它在分类过程中未考虑不平衡数据集的特点，还需要进一步的改进。因而基于 Boosting 方法也发展出了很多处理类别不平衡问题的算法，比如 1999 年 Shawe-Taylor 和 Grigoris Karakoulas John 提出了 AdaUBoost 算法[16]、2001 年 Joshi M V 和 Agarwal R C 等人提出了 RareBoost 算法[17]、Viola P 和 Jones M 提出了 AsymBoost 算法[18]、2003 年 Chawla N V 和 Lazarevic A 等人提出了 SMOTEBoost 算法[19]、2004 年 Guo H 和 Viktor H L 提出了 DataBoost 算法[20]、2012 年 Zhang X 和 Cheng C 提出了基于 Boosting 和级联模式的算法[21]等。此外，通过不同算法思想的集成，也可以对不平衡数据分类取得较好的效果，如周志华和刘胥影提出的代价敏感神经网络与分类器集成相结合的方法[22]。

2、代价敏感学习的基本思想是赋予各个类别不同的错分代价，因此它能够

很好地解决不平衡数据分类问题，基于代价敏感学习算法也发展出了很多处理不平衡数据分类问题的算法，比如 1999 年 Fan W 等提出了 AdaCost[23] 、2002 年 Ting 等提出了样本加权方法[24]、2003 年 Zadrozny B 等提出了 Costing[25]等。可以通过代价信息补偿类别间的分布差异，使得小类的样本有更高的错分代价。

3、此外还可以通过单类学习（one class learning）技术仅对小类样本进行学习，此类方法主要包括 Manevitz L M 等提出的使用单类支持向量机进行文件分类[26]、方景龙和王万良等人提出的用于不平衡数据分类的 FE-SVDD 算法[27]等。另外还有通过应用特征选择方法解决不平衡分类问题，Wang J 和 You J 等人就提出了通过特征选择达到不平衡数据分类的效果[28]。

目前已有很多不平衡分类算法应用于实际问题中。1998 年 Kubat M 和 Holte R C 等人就对卫星图像进行分析，应用改进的分类方法对石油喷井进行自动监测[5]。2003 年 Cohen G 和 Hilario M 等人分析了药物治疗检测中不平衡数据的分类问题，并将改进分类方法应用于生物信息学方面[29]。2004 年 Phua C 和 Alahakoon D 等人在金融欺诈的模式识别中讨论了不平衡数据分类问题[30]，Zheng Z 和 Wu X 等人分析了文本分类的特点，并提出针对不平衡数据分类的新方法[31]。2005 年 Peng Y 和 Huang Q 等人应用不平衡分类算法对基因数据进行分析来对乳腺癌进行诊断[32]。2012 年边婧和彭新光等人提出了新的大规模数据分层预处理 LDSP 算法，用于处理入侵检测大规模不平衡数据集分类问题[33]，这些应用均取得了不错的效果。

尽管类别不平衡学习已经取得了很多研究成果，但仍有很多值得研究的问题。在不平衡数据分类中，降取样方法因其有效性和高效性获得了广泛应用，但是它在降采样的过程中放弃了多数样本，没有充分利用已有的样本信息。因此，还需进一步研究如何在保持它的高效性的情况下弥补降取样方法的不足。代价敏感学习虽然能够有效地提高小类的预测准确率，但在大多数情况下很难对真实的错分代价做出准确的估计。单类学习方法比较适用于噪声较多的场合，对于噪声较少数据集进行分类时，性能非常不稳定。特征选择方法的在文本类数据分类上比较有优势，但使用范围较局限[34]。随着基于集成学习方法的不断成熟，它的应用将更加广泛深入。